Homework I
Datamining
Computer Science
Indiana Universty
Bloomington, IN

M.M. Dalkilic

January 26, 2015

Introduction

This second homework is straightforward. The problems will help elucidate the material. Please remember to label the homework correctly and include your statement of ownership. There are plenty of opportunities for you to do more and, consequently, learn more. There are three problems with assigned proportionality of value.

Structure of the Report

For the structure of the two databases, include each as a separate appendix. For your aglomerative code, create a separate appendix. Grading will include performing queries on your database as well as inspecting the code and watching it run. We will ask for the source code as well.

Problems

1. (6%) Statistics, Random Variables, Expectation
   Assume you and a friend play a game: throw a pair of six-sided, fair dice (d_1, d_2) to decide who is more generous. If you throw a 9 $(d_1 + d_2)$ your friend buys you 3 cups of coffee–each cup of coffee costs \$3.33. If you throw any other number, you buy your friend one cookie–each cookie costs \$1.00.

   a. What is the sampe space Ω?

b. Define the random variable $X : \Omega \to \mathbb{R}$ for this problem that partitions the data based on a throw.

c. Define the function on the random variable $g(X)$ that maps a throw to winning or losing from X.

d. What is the mass function over X?

e. What is the mass function over $g(X)$?

f. What is $E[X]$?

g. What is $E[g(X)]$?

h. From question c. above, what is a winner more likely to have, coffee or cookies?

2. (47%) Multiple Linear Regression, Database, Parametric, Readings
For this problem, use R as both a modeling and visualization tool, and incorporate MySQL to hold data. I will provide a flat ASCII file that contains information about used vehicles [2]. In particular, a tiny sample of used automobiles are priced using a well-known automobile valuation and research company. Presuming the pricing is reasonably good (state our assumptions), we explore how linear models can be used to explore data. Here are the names of attribues (or features) with descriptions when necessary:

- **Price** There are multiple prices associated with a vehicle. This price reflects a vehicle in "excellent condition." Although not very meaningful–in a verifiable sense–the vehicles are less than one year old.
- **Mileage**
- **Make**
- **Model**
- **Trim** Identifies added options.
- **Type** The sort of use intended for the vehicle.
- **Cylinders**
- **Liters** The size of the engine (metric).
- **Doors**
- **Cruise** Boolean indicating presense of technology.
- **Sound** Boolean indicating presensen of technology.
- **Leather** Boolean indicating material of seats.

a. Read this paper, "A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression," by R. R. Hocking [1]. You will need to register at JSTOR–which is free–to read the paper. In this paper, the author lists uses of linear regression (Mallows). What are these? What is ridge regression?

b. Create an MySQL database that stores the data in a single table and then load the data. Here is the screen capture of the table for the attributes. Your table should be identical.

```
mysql> describe veh_pri;
+-----------+-----------------------+------+-----+---------+-------+
| Field     | Type                  | Null | Key | Default | Extra |
+-----------+-----------------------+------+-----+---------+-------+
| price     | decimal(8,2)          | YES  |     | NULL    |       |
| mileage   | mediumint(8) unsigned | YES  |     | NULL    |       |
| make      | varchar(20)           | YES  |     | NULL    |       |
| model     | varchar(20)           | YES  |     | NULL    |       |
| trim      | varchar(20)           | YES  |     | NULL    |       |
| type      | varchar(20)           | YES  |     | NULL    |       |
| cylinders | tinyint(3) unsigned   | YES  |     | NULL    |       |
| liters    | decimal(3,1)          | YES  |     | NULL    |       |
| doors     | tinyint(3) unsigned   | YES  |     | NULL    |       |
| cruise    | bit(1)                | YES  |     | NULL    |       |
| sound     | bit(1)                | YES  |     | NULL    |       |
| leather   | bit(1)                | YES  |     | NULL    |       |
+-----------+-----------------------+------+-----+---------+-------+
12 rows in set (0.00 sec)
```

You should have 804 tuples:

```
mysql> select count(*) from veh_pri;
+----------+
| count(*) |
+----------+
|      804 |
+----------+
1 row in set (0.02 sec)
```

You can now do some preliminary analytics. For example, how many different cylinder types are included in the table? What are the associated counts?

```
mysql> select cylinders, count(cylinders) from veh_pri group by cylinders;
+-----------+------------------+
| cylinders | count(cylinders) |
+-----------+------------------+
|         4 |              394 |
|         6 |              310 |
|         8 |              100 |
+-----------+------------------+
3 rows in set (0.00 sec)
```

c. Assume we want to know the best set of predictors for price, *i.e.*

$$price \quad = \quad \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{1}$$

where $x_1, x_2, \ldots, x_k \in \{mileage, cylinders, \ldots, leather\}$ for $1 \leq k \leq 11$.

Your task is to

i. Create histograms for each of the attributes and visualize.

ii. Do a correlation for each pair $\rho(x_i, x_j)$ and visualize.

iii. Build a table that uses the linear model function from R for every possible model. You The table should show which attributes are used. Report the p-value for the F statistic and the adjusted R^2. For each pair of models, m_i, m_j, use R's `anova(`m_i`,`m_j`)` function to compare, reporting relavent information.

iv. What is a p-value? What is R^2? How do you interpret each entry in the table of (p, R^2) values:

p	R^2
very small	very small
very large	very small
very large	very large
very small	very large

3. (47%) Clustering agglomerative algorithms
For this problem, you will implement the agglomerative clustering algorithm. The data is found: `http://grouplens.org/datasets/movielens/`. You will use the MovieLens 1M dataset:

```
MovieLens 1M

1 million ratings from 6000 users on 4000 movies.
README.txt
ml-1m.zip
```

a. Download the three files and then design and build a database for them. Load them into a MySQL database. Create the three tables corresponding to the data:

movies, ratings , users

Decode a string into a tuple. For example, the ratings data is encoded as:

```
UserID::MovieID::Rating::Timestamp
```

where

- UserIDs range between 1 and 6040
- MovieIDs range between 1 and 3952
- Ratings are made on a 5-star scale (whole-star ratings only)
- Timestamp is represented in seconds since the epoch as returned by time(2)

Attributes would be named: userid, movieid, ratings, timestamp. The types are obvious. Do similar constructions for movies, ratings.

i. There are errors in the data; clean the database. What is the spurious data? How have you cleaned it? Does the change affect your algorithm? The document briefly discusses potential errors. What kind did you discover?

ii. Using SQL, do female and male computer scientists give similar ratings to movies that they both watch? What is the SQL that you used. What is your answer?

iii. Using SQL, who has rated more movies, men or women?

iv. Using SQL, which genre has the most ratings?

v. Implement an agglomerative clustering algorithm.

- Cluster on `age, profession`. What degree of similiarity exists for genre and ratings?
- Cluster on `genre, ratings`. What degree of similarity exists for age, sex, profession?
- Discuss the distance functions you used to cluster.
- Provide some measure of the quality of clustering.

4. (Extra 5%) LaTeX, bibtex
When you create a write-up of your homework, give the citations that are used in this homework and provide three others that are pertinent to you. Use the bibtex tool to generate citations.

References

[1] R. R. Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

[2] Shonda Kuiper. Introduction to multiple regression: How much is your car worth? *Journal of Statistics Education*, 16(3), 2008.